Icahn School of Medicine at Mount Sinai

# Data transformation

## Summarising and joining datasets

Joselyn Chavez

05/10/2022

# Let's recap

- **tibbles**
- **Import** and **export** tibbles
- **filter** data by row
- **select** columns (variables)
- **pivot longer** (columns to rows)
- **pivot wider** (rows to columns)

# summarise()

# Calculate the average age of sinai covid patients

```
# load libraries
library(tidyverse)
library(janitor)

# Import data
sinai_covid <- read_csv("sinai_covid.csv")

# clean names
sinai_covid <- sinai_covid %>%
  clean_names()
```

```
# summarise data
sinai_covid %>%
  summarise(mean(age))
```

```
## # A tibble: 1 × 1
##   `mean(age)`
##         <dbl>
## 1        61.3
```

Even better, assign a name to the result

```
sinai_covid %>%
  summarise(mean_age = mean(age))
```

```
## # A tibble: 1 × 1
##   mean_age
##      <dbl>
## 1     61.3
```

# Calculate more than one summarise

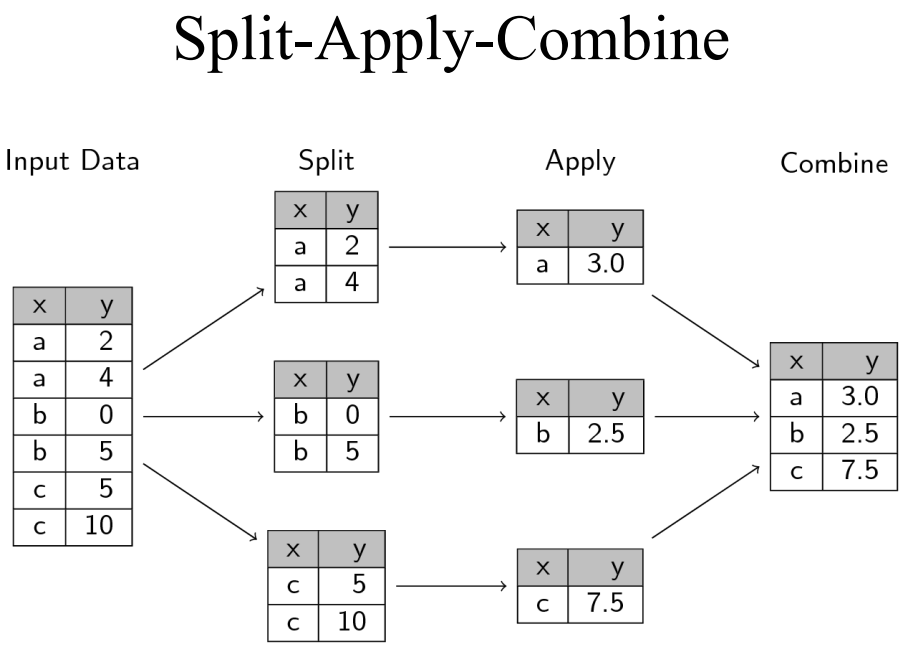
```
sinai_covid %>%
  summarise(mean_age = mean(age),
            sd_age = sd(age))
```

```
## # A tibble: 1 × 2
##   mean_age sd_age
##      <dbl>  <dbl>
## 1     61.3   16.9
```

# Your turn!

- Use summarise to calculate the median, max and min values of age

# How does summarise works?



# group_by()

Get the mean age and sd of sinai covid patients by sex

```
sinai_covid %>%
  group_by(sex) %>%
  summarise(mean_age = mean(age),
            sd_age = sd(age))
```

```
## # A tibble: 2 × 3
##   sex    mean_age sd_age
##   <chr>     <dbl>  <dbl>
## 1 FEMALE     62.2   18.0
## 2 MALE       60.6   16.0
```

# Your turn!

- Calculate the median age, and average bmi of sinai covid patients by facility

# We can use multiple variables to group by

```
sinai_covid %>%
  group_by(facility, sex) %>%
  summarise(mean_age = mean(age),
            sd_age = sd(age)) %>%
  head(4)
```

```
## `summarise()` has grouped output by 'facility'. You
## `.groups` argument.

## # A tibble: 4 × 4
## # Groups:   facility [2]
##   facility                    sex    mean_age sd_ag
##   <chr>                       <chr>     <dbl>  <dbl
## 1 MOUNT SINAI BI BROOKLYN     FEMALE     63.8   17.
## 2 MOUNT SINAI BI BROOKLYN     MALE       64.6   14.
## 3 MOUNT SINAI QUEENS HOSPITAL FEMALE     69.2   14.
## 4 MOUNT SINAI QUEENS HOSPITAL MALE       62.8   14.
```

# Let's practice!

- Find the mean, median and sd of the systolic_bp by smoking status and sex

# count()

For qualitative variables

```
sinai_covid %>%
  count(smoking_status)
```

```
## # A tibble: 3 × 2
##   smoking_status     n
##   <chr>          <int>
## 1 NEVER            341
## 2 QUIT             130
## 3 YES               29
```

# Excercise

- How many patients have chronic kidney disease?

- group and count

```
sinai_covid %>%
  group_by(facility) %>%
  count(smoking_status, diabetes)
```

```
## # A tibble: 27 × 4
## # Groups:   facility [5]
##    facility                    smoking_status diabe
##    <chr>                       <chr>             <d
##  1 MOUNT SINAI BI BROOKLYN     NEVER
##  2 MOUNT SINAI BI BROOKLYN     NEVER
##  3 MOUNT SINAI BI BROOKLYN     QUIT
##  4 MOUNT SINAI BI BROOKLYN     QUIT
##  5 MOUNT SINAI BI BROOKLYN     YES
##  6 MOUNT SINAI QUEENS HOSPITAL NEVER
##  7 MOUNT SINAI QUEENS HOSPITAL NEVER
##  8 MOUNT SINAI QUEENS HOSPITAL QUIT
##  9 MOUNT SINAI QUEENS HOSPITAL QUIT
## 10 MOUNT SINAI QUEENS HOSPITAL YES
## # … with 17 more rows
```

```r
sinai_covid %>%
  group_by(facility) %>%
  filter(diabetes == 1) %>%
  count(smoking_status,diabetes)
```

```
## # A tibble: 13 × 4
## # Groups:   facility [5]
##    facility                    smoking_status diabe
##    <chr>                       <chr>             <d
##  1 MOUNT SINAI BI BROOKLYN     NEVER
##  2 MOUNT SINAI BI BROOKLYN     QUIT
##  3 MOUNT SINAI BI BROOKLYN     YES
##  4 MOUNT SINAI QUEENS HOSPITAL NEVER
##  5 MOUNT SINAI QUEENS HOSPITAL QUIT
##  6 MOUNT SINAI ST. LUKE'S      NEVER
##  7 MOUNT SINAI ST. LUKE'S      QUIT
##  8 MOUNT SINAI WEST            NEVER
##  9 MOUNT SINAI WEST            QUIT
## 10 MOUNT SINAI WEST            YES
## 11 THE MOUNT SINAI HOSPITAL    NEVER
## 12 THE MOUNT SINAI HOSPITAL    QUIT
```

# join()

- Create tibble 1

```
tibble1 <- sinai_covid %>%
  filter(smoking_status %in% c("YES", "NEVER"
  count(smoking_status, name = "total")

tibble1
```

```
## # A tibble: 2 × 2
##   smoking_status total
##   <chr>          <int>
## 1 NEVER            341
## 2 YES               29
```

- Create tibble 2

```
tibble2 <- sinai_covid %>%
  filter(smoking_status %in% c("YES", "QUIT")
  group_by(smoking_status) %>%
  summarise(n_diabetes = sum(diabetes),
            n_obesity = sum(obesity))

tibble2
```

```
## # A tibble: 2 × 3
##   smoking_status n_diabetes n_obesity
##   <chr>                <dbl>     <dbl>
## 1 QUIT                    43        18
## 2 YES                      6         3
```

- Join tibbles
- Keep rows from left tibble

```
left_join(tibble1, tibble2)
```

```
## Joining, by = "smoking_status"
```

```
## # A tibble: 2 × 4
##   smoking_status total n_diabetes n_obesity
##   <chr>          <int>      <dbl>     <dbl>
## 1 NEVER            341         NA        NA
## 2 YES               29          6         3
```

- Keep rows from right tibble

```
right_join(tibble1, tibble2)
```

```
## Joining, by = "smoking_status"
```

```
## # A tibble: 2 × 4
##   smoking_status total n_diabetes n_obesity
##   <chr>          <int>      <dbl>     <dbl>
## 1 YES               29          6         3
## 2 QUIT              NA         43        18
```

- Keep rows in common

```
inner_join(tibble1, tibble2)
```

```
## Joining, by = "smoking_status"
```

```
## # A tibble: 1 × 4
##   smoking_status total n_diabetes n_obesity
##   <chr>          <int>      <dbl>     <dbl>
## 1 YES               29          6         3
```

- Keep all rows

```
full_join(tibble1, tibble2)
```

```
## Joining, by = "smoking_status"
```

```
## # A tibble: 3 × 4
##   smoking_status total n_diabetes n_obesity
##   <chr>          <int>      <dbl>     <dbl>
## 1 NEVER            341         NA        NA
## 2 YES               29          6         3
## 3 QUIT              NA         43        18
```

# Let's practice

- Count the number of patients by ethnicity, store the result in tibble 1
- Calculate the mean systolic_bp and mean diastolic_bp by ethnicity, store the result in tibble 2
- Join the tibbles

# Thanks!

Ilustration by Allison Horst